ML-lab Endsem Report

Josyula Venkata Srikar da24b006

November 2025

1 Models Used

Models used in the final prediction:

1. Preprocessing using PCA

2. Classification using KNN

Rejected Models:
Multiclass XGBoost. It was taking too long to train and wasn't giving good enough predictions. Train time was around 20 minutes and accuracy was 0.83.

2 Hyperparameter Tuning and Runtime

There were 2 hyper parameters, K value and the number of reduced dimensions for PCA. I tried different combinations and found the optimal ones as K = 7 and n = 100 for the validation dataset. But the difference is very small. Any n value in the approximate range 90-110 was giving 0.94 to 0.95 accuracy and K values from 5 to 20 were also giving similar scores. Keeping K = 1 gave a higher accuracy score but I think it might have been overfitting. Considering individual classes, 8 and 9 have lesser accuracy compared to other classes. Runtime was under a minute because I used matrix multiplications instead of calculating distance for each point separately. Initially, without matrix multiplication, it was taking around 3 minutes to run but after using matrix multiplications, it improved very much. The problem with this algorithm is there is no way to increase the depth or complexity so I am not able to increase the accuracy beyond 95.5%.

3 Thoughts

I think that to get good accuracy, we don't always need a very complex algorithm. Here, KNN itself is giving very good accuracy. Extending the binary classifier was not very easy. I used softmax probabilities for Multiclass XGB. It increased the train time very much as now we have to train more trees for each class. So it was not the most efficient way.